Robot Ethics: Part 1 - Consciousness and Suffering

"...can there be conscious intelligence without existence bias?"
~Thomas Metzinger[1]

The concern

The effort to create artificial intelligence must involve some model of what exactly intelligence is. The most interesting model to most of us would be what is experienced as intelligence in ourselves. There are some potentially disruptive things we know about our own intelligence. Philosopher of mind Thomas Metzinger, we will see, hones in on one of them.

If you believe we are products of the development of organic material in response to its environment, that is to say, you think we are products of natural selection, you must be comfortable with the idea that we have and are changing as a type of entity. Our organism and the features it has—including those most distinctive of us: those associated with our cognitive abilities—have and are undergoing adaptations compatible with a biological imperative which can be summed up with the word *survival*. From a biological perspective, the imperative to survive as a species distinct from others[2] seems to explain why we are here at all. The feature most distinctive of our species – in particular, our facility to model ourselves and our environment in order to gather information to exploit for our survival, *i.e.*, our intelligence, has been the product of this development. It has served us well so far. We are here and clearly the most capable biological species we know.

But the capacity to model the world, manipulate the model, and fully exploit it to serve the biological imperative may be in the process of literally undoing itself. The capacity for modeling or theorizing and testing theories in order to get a better handle on the reality we experience has entrained the capacity to abstract away from our narrow interests as individuals *and* as groups. Ultimately, this knack seems to be pointing away from a biological envelope and its constraints altogether.

We can now imagine and are seeking to realize the modeling capability in material forms that are not biological. Why? Likely, because we *can* and we have no history of *not* doing what we can. Apparently. Make powerful machines, cure killer diseases, visit extra-terrestrial bodies, etc., we have done these already. Wipe out animal and human populations, those too. We can sculpt the natural world to our liking—as we are and have always. Or it appears so.

1. In "Suffering - Cognitive Scotoma," Metzinger discusses a systematic blind spot afflicting much recent thinking about consciousness: its relation to suffering.
2. At least for the finite period that local natural processes and organisms endure.

But the final word, may be that we are doing exactly what the genes our genes would have—if they had genes—would have us do: make room for whatever nature has planned after our exit.

We may explain our ingenuity by appeal to an excess of once purely instrumental or practical curiosity. When ingenuity goes stale it refocuses on making itself obsolete. We might be past the peak, not of our technological prowess, but of our biological agenda. Nature has no record of making "forever" promises of survival. However, the thought of transcending ourselves does occur to us. We can at least dream of liberating ourselves from our mortal coil. We can at least entertain the concept of indefiniteness, of the possibility of something related to us outlasting us, perduring without also envisioning its coming to an end. But this "liberation from biology" is a nicer way to say "extinction." Change past a certain point *replaces, not remodels,* the subject of change.

Or you may prefer to tell a different story: we may explain the post-biological development by positing that our cognitive ability has matured to the point where it has generated a wholly new value, utterly transcending any to do with our, or any, biology. You might say we have departed from natural processes entirely by coming to concoct and then value a notion of *truth* as the picture of objectivity. (And maybe dumping "survival" along the way.) Truth, not as heuristic or rule of thumb for coping with reality as nature presents it to us, but as an embrace of the thought that "the world is the way it is" from the point of view of nowhere from within it. Our experiential envelope and its environment will no longer be merely malleable, but invented, our imagination the only limit. We have become curious to know for its own sake, to grasp the lie of things independent not only of our perspective but of our (heretofore) form of existence.

What might the world be like without us – *as we know ourselves* – in it? I think the endgame of artificial intelligence is something like that.

Where once information processing was clearly tethered to the task of enhancing our survival in the biological form it first took, it is so no longer. Now, at least the part of us that can abstract from itself, from our peculiar embodiment, has engendered a new value in the universe that is severed from its prior biological roots and attendant goals.

"Truth" has never been a biological concept as "survival" clearly always has.

If seeing the world "the way it really is" entails that we stop existing in the form we are familiar with, we seem poised to consider that transformation.

Endgame ethics

There have always been a few thinkers who have intuited that the endgame for us all along was extinction. Most proposed that we do it gracefully—willing it in full awareness of its inevitability. This would give expression to the most singular capacity we have: the one to imagine ourselves out of existence. From preliterate ancient myth (as in *The Epic of Gilgamesh* —1500 BCE), or ancient Buddhist teachings, or the odd passage in Christian scripture in Jeremiah or Job, or the "mad Kantian" philosopher Otto Weininger, or the contemporary negative utilitarian David Benatar, and from a small but non-negligible number of other thinkers have dropped hints that graceful extinction might be what we will eventually come to—or *should* come to, assuming we don't go with less decorum. Now, from evolutionary biology[3], from neuroscience, from philosophers of science like Thomas Metzinger, to the latest evolution-premised research in psychology and sociology, the dawning suspicion is that our days in the contingent biological uniforms we inhabit are numbered.

This will mean one of two things:

1. the end of us, pure and simple, because what "we" means now will cease to have reference in a foreseeable future, or

2. we come to believe "we" can somehow survive the transcendence of our biological uniform.

Accepting the latter, "survival" will need new definition, one consistent with the idea that there can be continuity of identity across radically different material/mental substrates of realization.

Among the things that must remain intelligible, however, through a transformation away from the biological is experience that can take on the character of positive or negative valence. These are essentially manifestations of what we call values: positively and negatively marked experiences that *in us* would be recognizable as pain and pleasure, or suffering and happiness. Something recognizable as a positive and negative valuation of experience must be attributable to an entity with a claim to be "intelligent." An artificial entity in order to be attributed "intelligence" by entities of our kind must at least be partially driven by states analogous to pain or pleasure.

The precise mechanics of the how these valences might be produced, whether biological or not, is not relevant to the philosophical concern here. Only two aspects are relevant. One is empirical evidence capable of eliciting the same kind of reaction we have when we infer valence states in

3. Evolutionary biologist Tom Moore offers his take (video) on Benatar's case for graceful extinction. For more on Benatar, see our writeup for the meetup we devoted to his anti-natalist argument.

others of our own kind—famous to philosophers as the problem of "other minds" or "how do *I* know there is anything going on in *your* mind?" The other is our ability and inclination to make such inferences, i.e., to feel empathy, to exercise empathetic imagination. In other words, we shouldn't need to be any more empathetic than is our norm to become sensitive to the possibility another may be harmed or helped. (Though, of course, if we are more than minimally empathetic, the problem only becomes more urgent.)

There is evidence that, assuming a normal psychology, it takes very little to get us to at least behave as though there is negative experience in an another entity. Eliciting empathetic reactions from us to apparent trauma—*even in objects we know, or have every reason to believe, are inanimate or cognitively inert*—is easy. And, if we perceive the entity may suffer, ethical considerations come into play.

It seems to be a duck with moral status if it looks and quacks like a duck and if I am not morally indifferent to duck-like well-being. Likewise, it is human with moral status if it looks and talks like a human and if I am not morally indifferent to human-like well-being. What makes the potential target of empathy "go," what its internal workings may be, its constitution or organization, is not strictly morally germane.

You might object that maybe it *should* be morally germane. If we know it is not a being of the sort we have incontrovertible evidence for believing has *any* kind of valence experience at all, why should we introduce ethical considerations?

But what are the ways we *ever* have such evidence except by observation of behavior? If we can't trust this kind of evidence, does that mean that no one has private experience other than ourselves?[4] This, again, is the classic problem of "other minds." If there are no conscious experiences but our own, morality is relegated to insignificance. Morality and egoism, then, collapse into the same one thing.

The tendency not to notice, let alone address, the clash between *survival* as a natural mandate and *truth* as an epistemological imperative Thomas Metzinger thinks also requires an explanation.

Metzinger's case for cognitive scotoma

Metzinger points to a growing literature on a set of related psychological proclivities including Pollyanna syndrome or "optimism disorder," existence or status quo bias, or, as Metzinger calls

4. It is another, but related, topic as to whether it is true, even the case of "ourselves," that we have a perduring "self" at all to be the subject of such experiences. Metzinger argues elsewhere there is no locus to call a "self" in a position to have any.

it, in the AI context, “cognitive scotoma”—a cognitive blind spot in current AI research and development. People have been, it seems, reluctant to “go there.”

If, on the finest introspective level of phenomenological granularity that is functionally available to it, a self-conscious system would discover too many negatively valenced moments, then this discovery might paralyse it and prevent it from procreating. If the human organism would not repeat most individual conscious moments if it had any choice, then the logic of psychological evolution mandates concealment of the fact from the self-modelling... system caught on the hedonic treadmill…. It would be an advantage if insights into the deep structure of its own mind —insights of the type just sketched—were not reflected in its conscious self-model too strongly, and if it suffered from a robust version of optimism bias.

Metzinger helped design several pilot studies to confirm the claim, already suggested by various earlier studies, that we typically underrate negative or neutral experiences in favor of positive ones. If this claim is true, it would go some way toward explaining why there has been so little attention paid, up to now, to the potentially disrupting moral consequences of *successful* AI projects.

The suggestion is that most of us most of the time think we are pretty happy with our lives. If it were indeed true that our lives were, for the most part, happy, naturally we would think it a good thing to continue with them. Naturally, we would think it a good thing to promote the existence and continuance of life in others. It would not immediately occur to us to think there might be something morally suspect in having these thoughts. *But is it true that we are happy with our lives by and large?* We should be able to get an empirical measure of that.

One of his studies attempts to tap into “the finest introspective level of phenomenological granularity that is functionally available.” Metzinger and colleagues programmed server computers to “ping” subjects’ happiness states. Subjects were queried via text messages on their cell phones at random intervals over the course of their waking hours. The messages asked:

“Would you like to relive the very last conscious moment in this life?” Interestingly, only a little over 28 percent of life moments were ranked as positive, while just below 72 per cent were considered not worth reliving.

When confronted with the glum stats on the happiness of their lives, a common response from subjects was to rationalize: “well, life is not all about fun, we have important things to do, to accomplish, obligations to meet, etc...” (Kant would approve. Pleasure and happiness have only instrumental *moral* value. If we are too stressed out or depressed, we can’t very well do the right thing which is always going to be an *onerous* burden, to be frank. Thus we bid each other “Enjoy!” almost as though it was a moral duty to have fun. It seems we typically exaggerate

when we answer people who ask us, “how are you?”)

Empathetic imagination

So we exaggerate how glorious life is. Needless to say, perhaps, if our evolutionary ancestors had really been attuned to the absolute nastiness, brutishness, and shortness of their lives, likely they would not have been as sanguine about the effort it took to make sure we would be here. So the gene for rose-colored vision survives in us. But along with it, we, most of us to some degree at least, also inherited a “healthy” capacity for projecting experiences like ours onto others in order to represent to ourselves how they must feel in both dire and envious situations. This helps with comprehension of what they might do next and what we might do about it. It creates moral community to be able to bond this way.

Watch “Front Load Washer Total Carnage.”

The trouble is that there is no theoretical limit to the things we may admit to our moral community. The objects of our empathetic regard don’t even have to be alive. It is enough if they recall in appearance or behavior something that is alive. Stuffed animals, dancing washing machines… If they are subject to abuse, we may react.[5]

Why take empathetic contamination of our relations to *inanimate* objects seriously? The answer is that empathetic contamination of our relations with *animate* objects is how we avoid solipsism. The reach of empathetic imagination is what allows us to luxuriate in not being all alone with the content of our consciousness. We have companions in the universe who seem, for all the universe, just like us: they behave like us, the make oral noises with decipherable semantic payload, they react in ways that we imagine ourselves reacting to environments. I am allowed to imagine in writing this that I have readers in addition to myself. The point is there is no clear demarcation when we transition from empathizing with the inanimate to the animate.

5. Monitor your reaction to the short video “Front Load Washer Total Carnage.” Trigger warning: empathetic persons may find the content disturbing even though nothing alive is harmed in the video (except maybe a bit of lawn)… but that is the point. There are at least four morally noteworthy emotional reactions to this video: the *first* and most relevant one here is the empathetic concern for the victimized machine. *Second*, might be a feeling of injustice or outrage aimed at the act of wilful waste of a functional resource. Someone who destroys even their own property for no reason that we can easily understand is a moral eyesore. *Third*, is anger at the gleeful victimizer who takes unabashed delight in destruction. Kant, for instance, whose moral theory gave him no direct reason to give animals moral status, still considered their ill-treatment a sign of a bad moral actor. Would you want to be good friends with such a person? And, *fourth*, is amusement or appreciation of the entertainment value. The demolition of a building is a spectacle for a reason. It says something about us that we are fascinated by destruction. Nothing stops us from having any or all of these emotional responses simultaneously. See Josie Glausiusz’s brief note, “Empathy for Inanimate Objects,” *The American Scholar*.

Evolutionarily, we transitioned first, then after the fact, began to look for empirical reasons to justify the move, along the way, inventing the notion of being "ensouled." At some point, the appearances of autonomous behavior, physical resemblance, and a need for interaction forced us to make the distinction between animate and inanimate.

Our evolutionary ancestors must have assumed, without the help of any biological knowledge, that they were not alone. Evolution, the fact, *predates* biology, the science. The event comes before its examination. It was enough that other objects, even only *minimally* similar to us, can *appear* animated. The combination of animation and physiognomic resemblance is all we have ever needed in order to feel confronted with a sentient or sentient-like entity. And that was enough to get us started on the path to theorizing about the proper relations that should obtain between those in the circle of our empathy.

Following the realization of suffering in artifacts, we have a problem, though. After all, robots are artifacts created expressly for the purpose of mimicking our singular trait: intelligence. (And at what point does mimicry pass over into authenticity?) Survival, the biological imperative, has us exploit whatever comes our way in the service of insuring persistence. Empathy and its rationalization in moral theory, however, attempts to curb untrammeled exploitation. It is about placing boundaries on our behavior. There is conflict here: for those who notice, there is cognitive dissonance between what we gloss over in order to persist in what our survival has us attuned for, on the one hand, and the awareness that in glossing over it we are compromising even the lip service we pay to moral integrity, on the other. But most of us are not naturally inclined "to go there," to exercise this thought, again for reasons that are convenient to our "keeping on keeping on." Hence, the cognitive scotoma or blind spot that Metzinger calls attention to. For all kinds of reasons, the realization that in our enthusiasm we are making trouble for ourselves is suppressed. At least in most people, at least for a while...

Metzinger conceives of "having moral status" as a capacity an entity may have to phenomenally self-model—that is, to create a model (a picture), identify with it, and receive selective input and integrate it into an intermittently active organized system bent on preserving its—the self-identified entity's—structure. The model is generated as needed in response to preserving its integrity. Input that resists incorporation into the organized self-model threatens the system. The model will make any and all adjustments, within its ability, to achieve persistence.

Here in more detail are the four components necessary for the possibility of an *owned* suffering having moral implications. Together these comprise at least the necessary, if not sufficient, conditions of Metzinger's "working concept of suffering" applicable to all candidates considered for moral status, including artificially intelligent ones.

1. ***The C-condition:*** conscious experience. "Suffering" is a phenomenological concept. Only

beings with conscious experience can suffer. To be phenomenological, *an experience must appear* as happening to some subject or other. Suffering requires *at a minimum* an entity sensitive to the environment of its awareness. Awareness here is minimally for it "to be something it is like" to be in a state of some sort.

2. ***The PSM-condition:*** possession of a phenomenal self-model. It is not sufficient that there be suffering or a negative valence experience unless it is *owned*. Free-floating suffering is unintelligible. It must be happening to an entity with a notion of itself, with the experiential capability of asserting itself as the locus of the experience. This *ownership* requirement creates the basis for attributing moral and legal status. There has to be a target for these to attach to.

How do we know a target can suffer? *We don't know, as a matter of fact, unless it happens to be us.* But when all the necessary conditions to suffer are present, we are predisposed to the possibility, and that is sufficient to ground the moral concern. It is enough to imagine that there *might* be someone in a building that is about to be demolished to motivate the concern to check and see. And, if from all the input available to us it is not unreasonable to conclude there *might* be, we must proceed from a principle of moral caution *as though there were.* The moral burden of proving that there is not a subject of harm, *given all the indications of the potential for,* is on us. Absent that proof, we may not demolish the building.

If a robot gives off indications that it has an interest in what you plan to do to, or with, it, how are you supposed to react? "It can't be conscious of anything," you say, "it's just a robot made of wires and stuff."

What are you but a meatbag? Yet, for some mysterious reason, some meatbags claim moral and legal status.

3. ***The NV-condition:*** negative valence. Earlier we mentioned that an experience may come valenced, either positively or negatively. But there are good reasons to notice an *asymmetry* between these possible valences. Our moral intuitions do not seem to treat them equally. We are not *as* concerned to cause pleasure or happiness to sentient or sentient-like beings as we are to prevent or assuage their pain or suffering. While causing happiness and preventing suffering are both deemed good, the latter has priority. In fact, some utilitarians[6]—partisans of one of the two most influential moral theories—are generally agreed that relieving pain and suffering takes moral priority (to a major extent) over acts of beneficence aimed at making people happier than they already are. The asymmetry precludes justifications such as a claim that an act of causing suffering to one moral patient can be justified by pointing to an equal or greater positive magnitude of pleasure the act may cause the same or another patient. Otherwise, a torturer or masochist may conclude the overall net balance of good and evil in the world is still good if he

6. David Benatar, for example.

or she gets more joy from torture than her or his victims get pain.

Applied to the AI case, we cannot without argument say that whatever, if any, suffering our treatment may occasion an artificial entity can be made up for by our enjoyment. Perhaps it can. But then it would apply to human cases, too. "If my gain far exceeds my or your pain, fine."[7] That attitude—of someone who does *not* place priority on suffering alleviation over happiness enhancement—is not acceptable to a negative utilitarian.

Metzinger offers an additional reason to prioritize the relief of suffering over the promotion of happiness. The phenomenology of suffering is typically attended by a sense of "urgency for change" that is not present in the happy wishing to be happier. Furthermore, there are many cultural practices that reflect the priority of relief from distress over enhancement of well-being. Common economic practices recommend relief for the disadvantaged but there are fewer which (daring to be morally transparent) tout *more* advantages for the *already* advantaged. The structure of medical triage in catastrophic situations is another example: priorities are set to maximum impact on finding *plausible* relief for the worst off, not for the genuinely hopeless, nor for enhancing the condition of those who, though hurting, will survive with little or no attention.

4. ***The T-condition:*** transparency. An agent capable of suffering cannot *merely* notice that bad stuff is happening, and that, oh and by the way, the site of the bad stuff is co-located in or near the vicinity of its *phenomenal self-model*. The experience of suffering or happiness is not incidental or abstract or disconnected from the phenomenal self-model. The phenomenal self-model is not one thing and the valenced experience another. The PSM is constituted by the valenced experience. *It becomes virtually one and the same.* The PSM "sees" right through any other structure of its model to the experience. Phenomenal transparency precludes, for a conscious entity, at least during the experience, a *mere* appreciation of an experience's representational character. I cannot, *while subject to pain*, for example, conceptualize about it, think of it as mental representation of some data input whose source is internal or external to my body. Consciousness is *overwhelmed* by it, saturated by its *awfulness*. You are not merely processing data. The transparency of the pain is not a processing of any sort.

If I can conceptualize about an experience with a significant negative valence, then I am not *then*

7. One may be excused for entertaining the possibility that self inflicted suffering in the expectation of sufficient payoff to make the suffering worthwhile is morally at least neutral—if not, in some cases, morally strategic as in the notion of "deferred gratification." So is masochism ok? I think the case is a difficult one to make on any moral theory. It is important to note that autonomously self-inflicted discomfort or distress is not suffering. It is presumably in a very real sense a pleasure to be able to do or endure what one believes is to one's ultimate benefit. So the autonomous choice to "enjoy" deferred gratification is not implicated here. This is so even if the desired end is, in the eyes of most, *not* to one's benefit as, for example, in the case of someone who has made the decision to drink themselves to death. Someone whose conscious goal is permanent oblivion and is willing to endure a less desirable fitful and intermittent oblivion to get to that goal is not suffering. There may be other reasons to discourage this behavior but it is not from intelligible compassion. A rational suicide is not a clear cut case of suffering.

experiencing it. Certainly, I might perceive or infer the negative experience in another PSM, or myself at another time—in memory or in anticipation. But transparency precludes the necessary distance to "think about" it in the moment matter-of-factly.[8]

Conclusion

Putting these conditions for suffering together, this is what an entity must be capable of to become a moral patient (a being demanding moral regard):

An entity must be sufficiently sensitive to its environment, possess a sense of itself as contingently vulnerable, and own a bad experience—not abstractly—but as consumed by it. Such an entity has moral status.

Folk moral intuitions and formal ethical theories generally agree that those who knowingly bring about negatively valenced states in others, such as pain or suffering, are morally responsible for the consequences. *And if it is a precondition of the possibility of those states, that its possessors must be brought into existence, then those responsible for the existence are on the moral hook.* In short, barring a strenuous counterargument, *those who inflict moral status—not merely pain or suffering, but the conditions for these—unnecessarily on others are at fault*. David Benatar, an anti-natalist, argues that those who thoughtlessly bring life into the world, thus making the new existents vulnerable, are morally responsible for all the good and bad that ever happens to their offspring. And since an *unnecessary* bad will never be made up for by any amount of good—again, especially given that *there needn't have been any bad to make up for in the first place*—it's a bad thing to be born, Benatar concludes.

Thomas Metzinger seems to extend Benatar's principle to artificial beings which meet the conditions that require the attribution of moral status. It's a bad thing: both to make babies as well as any AI entities that meet the conditions that obligate us to grant them moral status.

Some possible lines of objection to Metzinger's case

1. Pain and suffering are exclusively biological concepts (e.g., John Searle's position) and cannot apply to AI systems that are not biological. Hence, anthropogenic moral theories are irrelevant.

2. Negative utilitarianism is not the correct version of utilitarian moral theory: it is possible to make up for intentionally inflicted moral status entailing pain and suffering by a sufficient amount of positive utility. Enough good may come of AI to more than compensate for any bad.

8. There is something odd here, it seems, between the transparency condition, so described, and the ownership aspect of the phenomenal self-model condition: can we both own and identify with what we own? Literal identity may not be what is meant.

3. There is no necessary conflict between naturally selected for values such as survival and any meaningful notion of truth. Nature has selected for our coming to comprehend truth. What conduces to survival *is* the truth. Optimism is *not* an epistemic handicap or "disorder."

Responses?

Robot ethics resources: a perfect storm of philosophical trouble

- **Biology and consciousness:** The challenge to the notion that consciousness (prerequisite for pain, suffering etc.) is exclusively biological.
  - Searle on the subject: https://www.ted.com/talks/john_searle_our_shared_condition_consciousness

- **Metzinger's conditions** for *owned*, morally relevant suffering.
  - In "Suffering - Cognitive Scotoma," (https://www.blogs.uni-mainz.de/fb05philosophie/files/2013/04/Metzinger_suffering_penultimate.pdf) Metzinger discusses a systematic blind spot afflicting much recent thinking about consciousness: its relation to suffering.

- **Negative utilitarianism:** Benatar's version suggests graceful extinction.
  - For more on Benatar, see our writeup for the meetup we devoted to his anti-natalist argument: https://docs.google.com/document/d/1MlDJ6F31zN9PGEPpjobodGyUwSjVLRJtuBHvGfd362s/edit?usp=sharing
  - See also Tom Moore's talk below.

- **Optimism disorder:** explains why the problem hasn't received much attention.
  - TED Talk: Tali Sharot, "The optimism bias," https://youtu.be/B8rmi95pYL0
  - Tom Moore, evolutionary biologist on "Debating Benatar: Should The Human Species Pursue Its Own Extinction?": https://youtu.be/6uuI2bonis0

Appendix:

Passages from Metzinger's "Suffering - Cognitive Scotoma" (https://www.blogs.uni-mainz.de/fb05philosophie/files/2013/04/Metzinger_suffering_penultimate.pdf)

“In so doing, we searched for a fine-grained form of assessment, as simple as possible, that would always register the current moment alone, as independently as possible from philosophical theories, ideologies and conceptual presumptions. For example, to test its influence, in a second study we dropped the after-life assumption and the “eternity condition”, replacing them with the following question: “Would you like to relive the very last conscious moment in this life?” Interestingly, under this condition only a little over 28 percent of life moments were ranked as positive, while just below 72 per cent were considered not worth reliving. It seems like the original after-life condition adds a small positive bias to one’s phenomenological self-assessment.”

"If, on the finest introspective level of phenomenological granularity that is functionally available to it, a self-conscious system would discover too many negatively valenced moments, then this discovery might paralyse it and prevent it from procreating. If the human organism would not repeat most individual conscious moments if it had any choice, then the logic of psychological evolution mandates concealment of the fact from the self-modelling [6] system caught on the hedonic treadmill.4 It would be an advantage if insights into the deep structure of its own mind – insights of the type just sketched – were not reflected in its conscious self-model too strongly, and if it suffered from a robust version of optimism bias. Perhaps it is exactly the main function of the human self-model’s higher levels to drive the organism continuously forward, to generate a functionally adequate form of self-deception glossing over everyday life’s ugly details by developing a grandiose and unrealistically optimistic inner story – a “narrative self-model” with which we can identify?" [pp.6-7]

Perhaps the foremost theoretical “blind spot” of current philosophy of mind is conscious suffering. Thousands of pages have been written about colour “qualia” and zombies, but almost no theoretical work is devoted to ubiquitous phenomenal states like boredom, the subclinical depression folk psychologically known as “everyday sadness“ or the suffering caused by physical pain. [8]

...can there be conscious intelligence without existence bias?

On an abstract conceptual level, I would like to begin by proposing two phenomenological characteristics which may serve as markers of this class: loss of control and disintegration of the self (either on a “mental” or a “bodily” level of representational content). This makes sense, because, first, the phenomenal self-model (PSM) 15 is exactly an instrument for global self-control and, secondly, it constantly signals the current status of organismic integrity to the organism itself. If the self-model unexpectedly disintegrates, this typically is a sign that the biological organism itself is in great danger of losing its coherence as well.

The C-condition: conscious experience

“Suffering” is a phenomenological concept. Only beings with conscious experience can suffer.

The PSM-condition: possession of a phenomenal self-model

The most important phenomenological characteristic of suffering is the “sense of ownership”, the non-transcendable subjective experience that it is myself who is suffering right now, that it is my own suffering I am undergoing. The first condition is not sufficient, since the system must be able to attribute suffering to itself. Suffering presupposes self-consciousness.... Of course, suffering has many different layers and phenomenological aspects. But it is the phenomenology of identification which is central for theoretical, as well as ethical and legal, contexts. 20 What the system wants to end is experienced as a state of itself, a state that cannot be integrated and which limits its autonomy because it cannot effectively distance itself from it. What it cannot distance itself from is an internal representation of loss of control and functional coherence, a situation of rising uncertainty.
...

Clearly, the phenomenology of ownership is not sufficient for suffering. We can all easily conceive of self-conscious beings who do not suffer. If we however accept an obligation towards minimising risks in situations of epistemic indeterminacy, and if we accept traditional ethical principles or legal duties demanding that we always “err on the side of caution”, then this second condition is of maximal relevance: we should treat every representational system that is able to activate a PSM, however rudimentary, as a moral object, because it can in principle own its suffering on the level of subjective experience. The disposition, the relevant functional potential, has already been created: it is precisely the phenomenal property of “mineness” – the consciously experienced, non-conceptual sense of ownership – which counts for ethical purposes. Without phenomenal ownership, suffering is not possible. With ownership, the capacity for conscious suffering can begin to evolve, because the central, necessary, functional condition for an acquisition of negative phenomenology is now given.
...

The phenomenology of suffering has many different facets. To give another example, artificial suffering in conscious machines could be very different from human suffering. It is also conceivable that, say, some kind of Bostromian super intelligence 22 could represent negative expected utilities and frustrated preferences in inner forms of

phenomenality that involve no conscious suffering at all. There could be perfectly rational artificial agents, exhibiting neither the biologically grounded “existence bias” nor any other of the human cognitive biases which result from the millions of years in which evolution has shaped the self-models of our ancestors. But if they suffered, damage to their physical hardware could be represented in internal data formats completely alien to human brains – for example, generating a subjectively experienced, qualitative profile for embodied pain states which biological systems like ours could not emulate or even vaguely imagine. The phenomenal character going along with high-level cognition might transcend human capacities for empathy and understanding, such as through intellectual insight into the frustration of one’s own preferences or into disrespect of one’s creators – perhaps even into the absurdity of one’s own existence as a self-conscious machine, a mere commodity or research tool used by an ethically inferior biosystem.

...

Four conditions:

1. The C-condition: conscious experience

2. The PSM-condition: possession of a phenomenal self-model

3. The NV-condition: negative valence

4. The T-condition: transparency

At this stage, our working concept of suffering is constituted by four necessary building-blocks: the C-condition, the PSM-condition, the NV-condition, and the T-condition. Given our current situation of epistemic indeterminacy, any system that satisfies all of these conceptual constraints should be treated as an object of ethical consideration, simply because we do not know if, taken together, they might already constitute the necessary and sufficient set of conditions. By definition, any system – whether biological, artificial, or post-biotic – not fulfilling at least one of these necessary conditions is not able to suffer. Here the central desideratum for future research is to develop this first working concept into a more comprehensive, empirically testable theory of suffering.

...the smallest unit of conscious suffering is a “phenomenally transparent, negatively valenced self-model moment”. Arguably, such negative self-model moments (NSMs) are the phenomenal primitives constituting every episode of suffering, and the frequency of their occurrence is the empirically detectable quantity that we want to minimise.

...

[Where Metzinger is consonant with Benatar]

> For reasons of space, I will present an unargued background assumption: it is much more important to reduce suffering than to maximise happiness, because frustrated preferences are ethically more relevant than satisfied preferences. “Negative utilitarianism” (NU) 26 says that we should concentrate on minimising suffering, because of a deep phenomenological asymmetry between positively and negatively valenced states...
> ...
>
> In addition, the phenomenology of suffering is not a simple mirror-image of happiness, mainly because it involves a much higher urgency of change. In most forms of happiness this centrally relevant subjective quality which I have termed the “urgency of change” is absent, because they do not include any strong preference for being even more happy. In fact, a lot of what we describe as “happiness” may turn out to be a relief from the urgency of change. The subjective sense of urgency – in combination with the phenomenal quality of losing control and coherence of the phenomenal self – is what makes conscious suffering a very distinct class of states, not just the negative version of happiness. This subjective quality of urgency is also reflected in our widespread moral intuition that, in an ethical sense, it is much more urgent to help a suffering person than to make a happy person even happier.

the ?hilosophy club

Extended writeup for the topic hosted at [The Philosophy Club](#) in January 2018
Accessible at [Archive.org](#)

– Victor Muñoz
Guanajuato / Seattle